На правах рукописи

Акберова Наталья Ивановна

СИММЕТРИЧНЫЕ СТРУКТУРЫ В НУКЛЕОТИДНЫХ ПОСЛЕДОВАТЕЛЬНОСТЯХ САЙТОВ ИНИЦИАЦИИ РЕПЛИКАЦИИ ДНК ПРОКАРИОТ

Специальность: 03 00 04 - биохимия

АВТОРЕФЕРАТ

диссертации на соискание ученой степени

кандидата биологических наук

Казань - 1999

Работа выполнена в лаборатории биохимии нуклеиновых кислот Казанского государственного университета

Научные руководители: доктор биологических наук, профессор В.Г.Винтер
кандидат физ.-мат. наук, с.н.с. А.Ю.Леонтьев

Официальные оппоненты: доктор биологических наук, профессор В.И.Чиков
доктор медицинских наук профессор В.В.Семенов

Ведущая организация: Институт биологии гена РАН, г.Москва

Защита состоится «___»_____1999 г. _____ час. на заседании диссертационного Совета К.053.29.19. при Казанском государственном университете, 420008, г. Казань, ул. Кремлевская, 18.

С диссертацией можно ознакомиться в библиотеке Казанского государственного университета

Автореферат разослан «___»_____1999 г.

Ученый секретарь диссертационного совета,
кандидат биологических наук А.Н.Аскарова

Актуальность темы. Репликация ДНК, один из фундаментальных процессов, отличающий живые существа от неживой природы, контролируется на уровне инициации. Инициация репликации многостадийный процесс, требующий участия многих белков, происходит в определенных участках хромосомы, которые получили название сайты инициации репликации или «origin», «ori».

В литературе неоднократно отмечалось, что понимание структуры ori - ключ к управлению инициацией репликации ДНК, в том числе патогенных организмов [Huberman J.A., 1995]. По мнению большинства исследователей, выяснение природы репликаторов эукариот признано одной из самых актуальных проблем современной молекулярной биологии [Abstracts of CSH Meeting, 1997].

Кроме экспериментальных методов исследования процессов, в которых принимают участие определенные участки ДНК (таких как инициация репликации ДНК), в последние годы получили широкое распространение компьютерные методы. Целью этих методов, в частности, является теоретическое изучение строения первичной структуры ДНК, что очень важно для понимания её роли в исследуемом процессе. Такие исследования являются мощным инструментом, позволяющим сократить число дорогостоящих экспериментов и спланировать дальнейшие практические исследования. Особенно актуальной является разработка компьютерных методов исследования ori потому, что практические исследования инициации репликации ДНК трудоемки, результаты, как правило, трудно интерпретируемы в силу сложности используемых экспериментальных методов и самого предмета исследования.

Для анализа нуклеотидных последовательностей предложено множество методов. Широко используются консенсусные и матричные методы [Bucher P., 1990; Karlin S, Brendel V., 1992; Quandt K. et al., 1995; Uberbacher E.C. et al, 1996; Chen Q.K. et al, 1997]. Применяются также методы, основанные на методологии статистики [АлександровА.А. и др., 1986; Соловьев В.В и др., 1987; Сприжицкий Ю.А и др., 1988; Франк-Каменецкий , 1990; Бородовский М.Ю., Певзнер П.А. , 1990 Pevzner P.A. et al , 1989; Mirkes E.M. et al, 1993], математической лингвистики [Brendel V. et al, 1986; Trifinov E.N., Brendel V., 1986; Searls D.B., 1997], методы информационного анализа генетических текстов [Гусев В.Д. и др., 1989; Martingale C., Konopka, A.K, 1996]. Интегрированные статистические методы, учитывающие сразу несколько характеристик текста, разработаны для компьютерного распознавания в геноме функциональных сайтов, таких как промоторы [Kel et al, 1995; Ponomarenko MP et al , 1997; Fickett, J.W., 1996].

Большинство методов анализа нуклеотидных последовательностей основаны на использовании статистического подхода, применение которого встречает ряд принципиальных трудностей связанных с самой природой статистики. В первую очередь это касается вида распределения вероятностей, на основании которого рассчитываются ожидаемые встречаемости тех или иных фрагментов генетического текста. В связи с этим становится актуальной разработка методов анализа нуклеотидных последовательностей функциональных сайтов, не опирающихся на априорные сведения о распределении плотности вероятностей. Такой нестатистической характеристикой нуклеотидных последовательностей может быть наличие в них различных симметричных структур. Известно, что такой элемент симметрии как повтор является характерной чертой регуляторных районов генома как прокариот, так и эукариот [Day GR, Blake RD, 1982; Boulikas T, 1996], что предполагает его важность для функции регуляторных областей генома [Pearson et al, 1996; Boulikas T, Kong CF, 1993; Samadashwily GM et al , 1997]. В 1992 г. был предложен контекстно-независимый подход для анализа структуры генетического текста [Леонтьев А.Ю , 1992], основанный на выявлении внутренней обобщенной симметрии текста, в котором в качестве характеристик генетического текста было предложено использовать наличие в них повторов различных типов симметрии.

Преимуществом рассмотрения симметрии первичной структуры ДНК и распределения симметричных фрагментов в качестве подхода для анализа генетических текстов является то, что в нем не требуется учета статистических характеристик текста.

В данной работе такой симметрийный подход применяется для исследования генетических текстов сайтов инициации репликации прокариот и предлагается метод построения симметрийных консенсусов для изучения первичной структуры ori.

Цель работы:

Целью данной работы явилось выявление внутренней симметрии сайтов инициации репликации на уровне первичной структуры ДНК, выяснение роли симметричных структур генетических текстов этих сайтов в процессе инициации репликации ДНК.

Были поставлены следующие задачи:

1. Выявить повторы всех возможных типов симметрии в генетических текстах сайтов инициации репликации ДНК прокариот.
2. Оценить эффективность симметрий как формальных признаков генетических текстов для функционально-значимой классификации изучаемых последовательностей.

3. Оценить эффективность предлагаемого симметрийного подхода для исследования структуры сайтов инициации репликации прокариот.
4. Построить формальную модель сайта инициации репликации ДНК прокариот, пригодную для компьютерного распознавания ori в геноме прокариот .

Научная новизна:

Впервые предлагается контекстно-независимый, нестатистический метод для исследования структуры сайта инициации репликации и для классификации нуклеотидных последовательностей с функцией ori, основанный на внутренней симметрии нуклеотидных последовательностей.

Разработанный в работе метод анализа нуклеотидных последовательностей позволил выявить повторы всех возможных типов симметрии в генетических текстах 13-ти прокариотических сайтов инициации репликации ДНК. Наряду с традиционно исследуемыми повторами (простые и комплементарные) выявлены структуры обычно не рассматриваемых типов симметрии - RY (пурин-пиримидин) и KM (амино-кето)

Показана значимость симметричных структур, выявленных в нуклеотидных последовательностях сайтов инициации репликации ДНК прокариот для функции ori.

Практическая ценность:

Предложен контекстно-независимый метод анализа нуклеотидных последовательностей, основанный на внутренней симметрии генетического текста функциональных сайтов, который может быть использован для классификации последовательностей .

Предложенный метод позволяет выявлять в последовательности значимые для функции структуры.

Разработан метод описания симметричной структуры достаточно протяженных участков генетического текста (100-1000 и более), пригодный для создания образов распознавания функционального сайта в геноме.

Публикации и апробация работы: Основные результаты диссертации опубликованы в работах [1-13], докладывались на международном симпозиуме BIOMATH-95 (Mathematical Modelling and Information Systems in Biology, Ecology and Medicine (Sofia, Bulgaria, August 23-27, 1995), на международной конференции по секвенированию и анализу генома (7th International Genome Sequencing and Analysis Conference (Hyatt Regency, Hilton Head, SC, USA, September 16-20, 1995), на симпозиуме и школе «Структура и функция генома» (2 nd EL.B.A, Foundation Course on Genome : NATO-ASI "Genome structure and function"(Marchiana Marina, Elba, Italy, June 13-23, 1996), были представлены на итоговых научных конференциях КГУ 1996, 1997, 1998 гг.

Структура и объем работы. Диссертация состоит из введения, 3 глав, заключения, списка литературы, приложения, изложена на 114 страницах, содержит 5 рисунков, 3 диаграммы и 10 таблиц. Список литературы включает 247 наименований, из которых 230 иностранных работ. Глава 1 посвящена, во-первых, обзору экспериментальных данных о строении сайтов инициации репликации ДНК прокариот, во-вторых, в ней описаны основные компьютерные методы анализа генетических текстов. В главе 2 описаны объекты и методы исследования, использованные в работе. В главе 3 представлены результаты работы с их обсуждением.

Объекты

В работе были исследованы генетические тексты ori репликации ДНК прокариот. Из имеющихся в банке нуклеотидных последовательностей Genbank 182 последовательностей, связанных с инициацией репликации прокариот, были отобраны только хромосомные ori, последовательности которых содержат информацию, достаточно полную для инициации репликации ДНК. Не включены в этот список последовательности с нечеткой аннотацией и идентичные (как в роде Shigella).

Таблица 1. Список изученных последовательностей

No	Организм	Индекс по GenBank	Длина (в нуклеотидах)	Условное обозначение
1	**Bacillus subtilis**	**v01490**	**486**	**bs**
2	**Escherichia coli**	**v00308**	**555**	**ec**
3	**Enterobacter aerogenes**	**j01576**	**556**	**ea**
4	**Klebsiella pneumoniae**	**j01744**	**554**	**kp**
5	**Erwinia carotovora**	**v00255**	**696**	**er**
6	**Pseudomonas aeruginosa**	**m30125**	**651**	**pa**
7	**Pseudomonas putida**	**m30126**	**651**	**pp**
8	**Shigella dysenteriae**	**x67657**	**450**	**sd**
9	**Salmonella typhimurium**	**j01808**	**552**	**st**
10	**Streptomyces coelicolor**	**m82836**	**921**	**sc**
11	**Vibrio harveyi**	**k00829**	**277**	**vh**
12	**Caulobacter crescentus**	**s43898**	**998**	**cc**
13	**plasmid R751**	**pAR757**	**870**	**pl**

Мы ограничились приведенными в Табл.1 тринадцатью примерами областей инициации репликации ДНК, куда включили наряду с хромосомными один плазмидный репликатор. Далее в работе для обозначения исследуемых последовательностей будем пользоваться условными обозначениями этой таблицы.

Методы

В настоящей работе для анализа нуклеотидных последовательностей сайтов инициации репликации прокариот был применен предложенный Леонтьевым А.Ю. [Леонтьев А.Ю , 1992] симметрийный подход и на его основе разработан контекстно-независимый, нестатистический метод построения образа ori прокариот, базирующийся на симметричных структурах, обнаруживаемых в их нуклеотидных последовательностях. Достоинством этого метода является использование контекстно-независимой характеристики нуклеотидных последовательностей ori прокариот - симметрии их генетических текстов.

1. Метод анализа симметрии генетического текста.

Генетический текст ДНК рассматривается как одномерная дискретная структура из точек, окрашенных в четыре цвета, соответствующих четырем типам нуклеотидов. Всего с учетом цветных и пространственных преобразований в генетическом тексте возможны 8 видов симметрии, базирующиеся на биохимически осмысленных гомоморфных преобразованиях алфавита ДНК - {a↔t, c↔g}, {a↔g, c↔t}, {a↔c, g↔t} и повторениях или инверсиях отдельных фрагментов [Леонтьев А.Ю , 1992]. Преобразование {a↔t, c↔g} соответствует комплементарным взаимодействиям; преобразование {a↔g, c↔t} преобразует пурин в пурин, а пиримидин - в пиримидин; третье преобразование {a↔c, g↔t} связывает основания с амино- и кето-группами, обращенными в большую бороздку ДНК. Вместе с инверсией текста эти преобразования определяют повторы 8-ми видов симметричных структур. Все эти структуры могут быть как разнесенными, так и тандемными. Этим 8-ми типам симметрии соответствуют строгие с математической точки зрения группы , что позволяет строить эффективные алгоритмы их распознавания в геноме.

Используемая в данной работе номенклатура симметричных структур, представлена в Таблице 2.

Таблица 2. Номенклатура симметричных структур в генетическом тексте, принятая в данной работе.

Типы симметрий	Обозначение симметрий	Ориентация повторов	
		Прямые (Dir)	Инвертированные (Inv)
Обычные	Cmn	aagct...aagct	aagct...tcgaa
Комплементарные	Cmpl	aagct...ttcga	aagct...agctt
Пурин или пиримидин	RY	aagct...ggatc	aagct...ctagg
Амино- или кето-	KM	aagct...cctag	aagct...gatcc

ПРИМЕЧАНИЕ : Пентануклеотид aagct выбран в качестве примера сайта, не имеющего внутренней симметрии

Алгоритм поиска симметрий реализован на IBM PC, основан на полном переборе вариантов, время поиска повторов всех возможных типов симметричных структур в нуклеотидной последовательности, длиной 500 нуклеотидов составляет 3 минуты .

Наша задача заключалась в том, чтобы выяснить, могут ли симметричные структуры быть использованы для полного и корректного решения задачи распознавания ori прокариот. С этой целью были изучены симметрийные паттерны исследуемых нуклеотидных последовательностей ori прокариот и построены образы этих функциональных сайтов для таксономических групп разных уровней.

Методика построения симметрийного образа состоит из следующих этапов :

1) для каждой исследуемой последовательности определялись все входящие в нее повторы всех типов симметричных структур, описанных в Таблице 2;

2) для выявления общих для исследуемых нуклеотидных последовательностей из всех найденных повторов всех типов симметрии для дальнейшего анализа отбираются только те, которые встречаются не менее, чем в двух последовательностях;

3) для построения "симметрийного консенсуса" (СК) выбирались лишь те симметричные структуры, которые расположены во всей рассматриваемой группе на фиксированных расстояниях друг от друга; таким образом под симметрийным консенсусом мы подразумеваем общий для группы последовательностей набор симметричных структур, находящихся во всех последовательностях данной группы на фиксированных расстояниях;

4) множество изучаемых последовательностей разбивается на подгруппы по признаку вхождения тех или иных симметричных структур; собственно процедура классификации заключается в построении дерева, в вершинах которого располагаются найденные группы последовательностей, а ребра образуют иерархию родства по рассматриваемым формальным признакам.

Для адаптации построенных в работе симметрийных консенсусов задачам компьютерного распознавания ori в геноме прокариот разработан формальный язык описания ori, в котором в качестве элементов словаря использованы симметричные структуры, обнаруженные в исследуемых последовательностях, эти элементы связаны отношениями «содержит», «слева» и «расстояние».

2. Метод оценки значимости выявленных элементов симметрии для функции ori.

Значимость выявленных в работе симметричных структур для функции инициации репликации у прокариот определялась:

1). при сравнении проведенной в работе, основанной на симметриях формальной классификации исследуемых последовательностей с содержательными классификациями; в качестве содержательных классификаций использовали таксономию организмов, которым принадлежат эти последовательности, и классификацию прокариотических ori, основанную на экспериментальных данных об их функциональной взаимозаменяемости;

2). при сравнении элементов построенных в работе симметрийных консенсусов с описанными в литературе структурными элементами ori прокариот, важными для функции.

Использованное в работе программное обеспечение разработано Леонтьевым А.Ю. в лаборатории биохимии нуклеиновых кислот КГУ, алгоритмы реализованы на IBM PC.

РЕЗУЛЬТАТЫ и ОБСУЖДЕНИЕ

Симметричные структуры в генетическом тексте последовательностей прокариотических ori

На первом этапе проведенного симметрийного анализа были выявлены повторы всех типов симметрии, входящие в каждую из 13-ти исследуемых последовательностей. Насыщенность исследуемых последовательностей повторами всех возможных типов симметрии оказалась очень высокой, и мы ограничились рассмотрением повторов длиной 6 нуклеотидов и выше.

Диаграмма 1. Встречаемость повторов различных типов симметрий (для повторов длиной 6 нуклеотидов)

Cmn - обычные; Cmpl - комплементарные; RY - "пурин-пиримидин"; KM - "амино-кето"

Во всех 13-ти исследуемых последовательностях были обнаружены повторы разной длины всех 8-ми типов симметрии. В изучаемых последовательностях простых и комплементарных повторов было обнаружено больше всего (относительно общего количества всех выявленных повторов). На диаграмме. 1 показаны соотношения типов выявленных симметричных структур в изученных последовательностях на примере повторов длиной 6 нуклеотидов. Такое же соотношение симметрий разных типов сохраняется и для повторов большей длины.

Анализ вхождения разных повторов в отдельные последовательности представлен на диаграмме 2 на примере повторов длиной в 6 нуклеотидов. На диаграмме показано, что соотношение встречаемости повторов различных типов симметрии неодинаково для индивидуальных последовательностей. Например, в {ea} выделяются инвертированные комплементарные повторы, а все остальные типы повторов встречаются примерно одинаково часто. Для {ea}, {kp}, {pa}, {pp}, {cc} выявлено большее, чем среднее по всей выборке относительное содержание RY- и KM-структур (более 10%), а в {vh}, {bs}, {sc} таких повторов встречается меньше (около 5%). Простых и комплементарных повторов больше всего (выше 20%) содержится в последовательностях {ec}, {sd}, {st}, {sc}, {vh}, {cc}, {pl}.

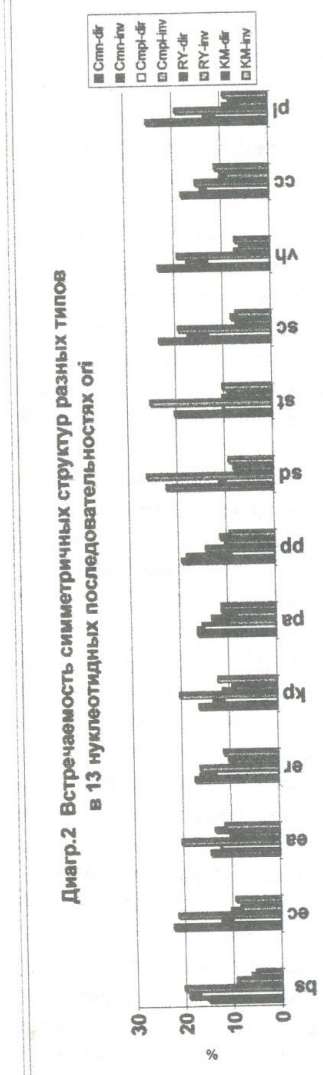
На диаграмме 3 показаны максимальные длины элементов обнаруженных повторов. В основном для всех последовательностей и всех типов симметрии длина элементов повтора лежит в диапазоне от 8 до 11 нуклеотидов. Выделяются максимальная длина элементов простых инвертированных (20 нуклеотидов) и комплементарных инвертированных повторов (14 нуклеотидов) в последовательности {bs}, в {pl} - максимальная длина элемента прямых простых повторов равна 18 нуклеотидам, инвертированных комплиментарных - 17 нуклеотидам, в {sc} - максимальная длина прямых простых повторов составляет 13 нуклеотидов. Самые длинные элементы повторов типа RY - 13 нуклеотидов - обнаружены в {pp}, самые длинные KM-повторы - 14 нуклеотидов - в {er}.

Таким образом, в нуклеотидных последовательностях изученных сайтов инициации репликации прокариот обнаружены повторы всех возможных 8-ми типов симметрии: в исследуемых последовательностях сайтов инициации репликации выявлены повторы как традиционно исследуемых типов симметрии (обычные и комплементарные повторы), так и повторы необычных RY (пурин - пиримидин) и KM (амино- кето-) симметричных структур.

Повторы симметричных структур RY- и KM- типов встречаются в исследованных последовательностях примерно в два раза реже, чем традиционно рассматриваемые повторы. Максимальные длины элементов повторов всех исследованных типов симметрии лежат в диапазоне 8-12 нуклеотидов.

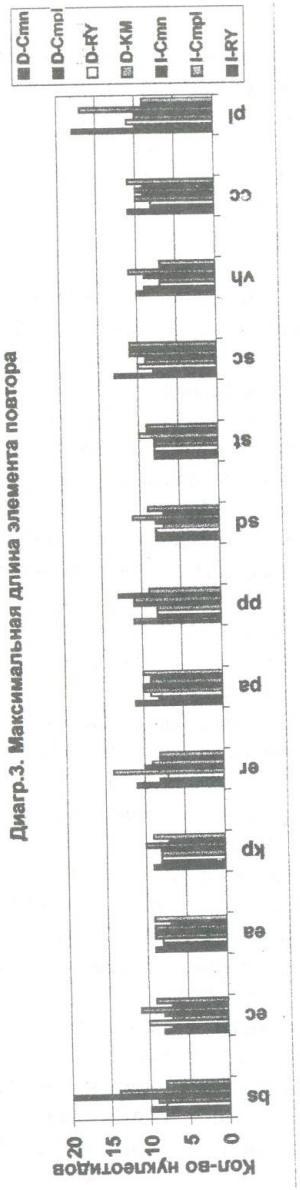
Классификация изучаемых последовательностей, основанная на вхождение в них симметричных структур.

Классификация изучаемых последовательностей ori, основанная на симметриях, была предпринята в предположении, что при совпадении



Диагр.2 Встречаемость симметричных структур разных типов в 13 нуклеотидных последовательностях ori

Обозначения последовательностей - как в Таблице 2; типов и ориентации повторов - как в Таблице 2.



Диагр.3. Максимальная длина элемента повтора

такой формальной классификации с содержательной (основанной на экспериментальных данных), признаки, по которым проводилась формальная классификация, можно будет считать значимыми для выполнения ori его функции. В качестве признаков для формальной классификации рассматривались наличие в последовательностях определенных симметричных структур и расстояние между элементами повтора.

Классификация множества генетических текстов по признаку вхождения симметричных структур

Для выявления общих симметричных структур для всех изучаемых последовательностей (или части из них) из всех обнаруженных повторов для дальнейшего анализа были отобраны только те, которые входят не менее, чем в две изучаемые последовательности. Следует отметить, что нам не удалось обнаружить ни одной симметричной структуры, которая объединяла бы все 13 последовательностей. Хотя во всех исследованных последовательностях были обнаружены сайты связывания DnaA белка, но они присутствуют в виде элементов повторов различных типов симметрии.

На стадии анализа, когда учитывается лишь вхождение определенных повторов в последовательности, изучаемое множество последовательностей подразделяется на подмножества. При учете расстояний между повторами классификация изучаемых последовательностей становится более детальной, и следующим этапом работы было объединение изучаемых последовательностей в группы по признаку присутствия определенных симметричных структур, элементы которых расположены в них на фиксированных расстояниях.

Симметричные структуры двух типов - прямой обычный и инвертированные комплементарные повторы - объединяют группы из восьми и семи последовательностей. Группа из шести последовательностей, выделяется по вхождению наряду с прямыми и инвертированными простыми повторами также и обратных КМ-повторов. Повторы, объединяющие пять последовательностей, представлены всеми возможными типами симметрии. Среди них выделяется группа { ec, ea, kp, sd, st }, которая объединяется не только вхождением различных симметричных структур шести типов, но также и фиксированными расстояниями между элементами этих структур. Прямой RY повтор cggatc и обратные комплементарные повторы gttatcc и gatccc объединяют в группу последовательности { ec, ea, kp, er, sd }, причем эти повторы расположены на строго фиксированных расстояниях во всех последовательностях этой группы.

Этот этап работы можно завершить следующим заключением:

Применение симметрийного подхода для анализа исследуемых последовательностей сайтов инициации репликации ДНК прокариот позволило выявить в них общие структуры, провести классификацию исследуемых последовательностей, основанную на таких формальных признаках как наличие определенных симметричных структур в последовательностях и расстоянии между элементами общих повторов.

Для формальной классификации наиболее значимыми оказались повторы следующих типов симметрии: прямые простые и прямые комплементарные, прямые RY повторы, инвертированные комплементарные, инвертированные RY и инвертированные КМ повторы. Это является свидетельством того, что «цветные» симметрии, наряду с традиционно рассматриваемыми, представляют важные свойства исследуемых последовательностей.

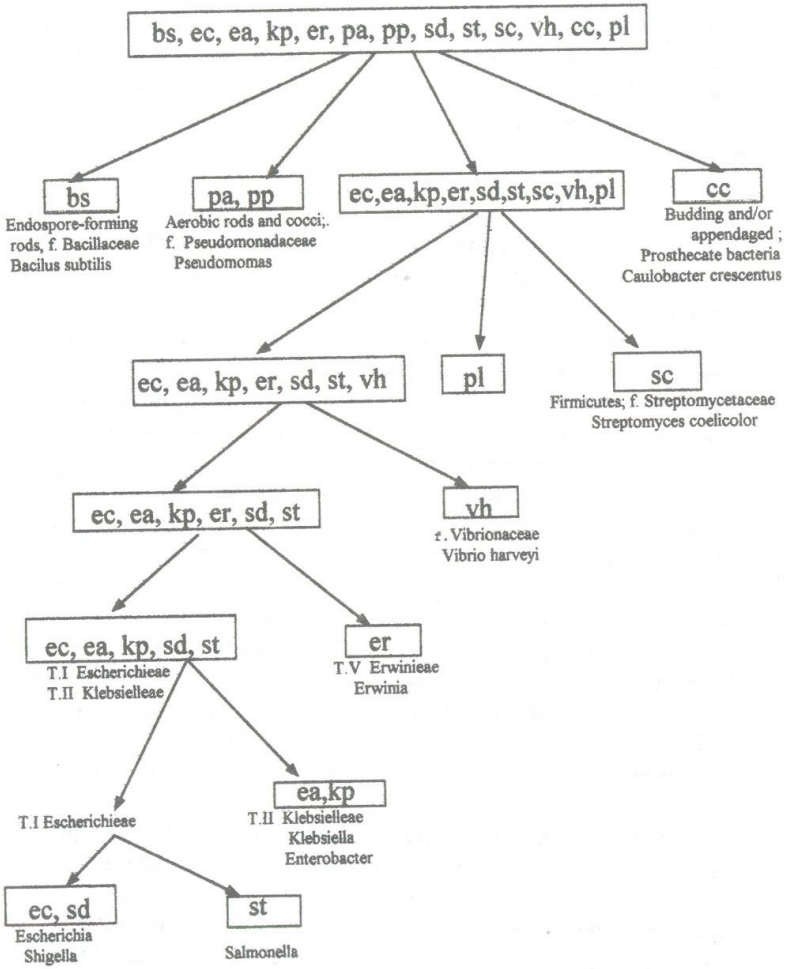
Соответствие основанной на симметриях формальной классификации изучаемых последовательностей классификации прокариотических ori, основанной на экспериментальных данных

На этом этапе работы оценивалась эффективность предлагаемого симметрийного подхода для решения задачи разбиения изучаемых последовательностей на группы по их формальным признакам с целью выявить ценность этих признаков для функционально значимой классификации генетических текстов ori. Суть процедуры классификации заключается в разбиении изучаемой группы последовательностей, осуществляющих одну и ту же функцию, на подгруппы по признаку вхождения в них симметричных структур. Полученное нами разбиение представлено в виде дерева родства изучаемых последовательностей на рисунке 1. Было обнаружено, что классификация последовательностей, основанная на вхождении симметрий, вполне соответствовала принятой филогенетической таксономии организмов, из геномов которых были взяты изучаемые последовательности. Это свидетельствует о том, что наличие симметричных структур в исследуемых текстах является содержательным признаком последовательностей ori, значимым для таксономии организмов.

Имеются экспериментальные данные о том, что ori репликации Bacillus subtilis, Caulobacter crescentus и Pseudomonas не могут функционировать в E. coli , а oriC неактивен в Bacillus subtilis, Caulobacter crescentus, Pseudomonas aeruginosa и Pseudomonas putida [Moriya S et al, 1992; Marczynski GT, Shapiro L, 1995; Smith DW et al, 1991]. Такая неспособность к «функциональной комплементации» ori этих бактерий отражает различия в их строении, поэтому принято относить их к разным

классам прокариотических ori [Yee TW, Smith DW, 1990; Ogasawara N, Yoshikawa H,1992; Zweiger, G et al, 1994].

Рис. 1 Дерево родства исследованных последовательностей, основанное на вхождении в них симметричных структур.



*Обозначение последовательностей как в Таблице.1; таксономическая классификация организмов взята из банка данных NCBI-Taxonomy Национального центра биотехнологической информации.

Проведенная в данной работе классификация последовательностей, основанная на вхождении симметричных структур, позволяет разделить последовательности, относящиеся к разным типам прокариотических ori и, наоборот, объединить в группу последовательности, для которых экспериментально была показана их функциональная взаимозаменяемость , что служит подтверждением значимости симметричных структур для функции инициации репликации ДНК.

Построение симметрийных консенсусов ori

В этом разделе оценивается пригодность симметричных структур в генетических текстах сайтов инициации репликации для описания внутреннего строения ori прокариот. Учитывалось не только наличие симметричных структур, но и их взаимное расположение (позиция слева-справа, расстояние как внутри повтора, так и между разными повторами). На основании вхождения в изучаемые последовательности одинаковых симметричных структур, элементы которых расположены на равных расстояниях, строились симметричные консенсусы (СК) для групп последовательностей.

Симметрийный консенсус для последовательностей группы E.coli

а) Симметрийный консенсус для {ec, ea, kp, er, sd, st, vh}

```
tgtggataa      ttatccaca
<********      ********>
|←———180———→|
```

б) Симметрийный консенсус для {ec, ea, kp, er, sd, st }

```
                    |←40-42→|
agatct     agatct          tgtggataa      ttatccaca
<*****     *****>          <********      ********>
----->     ----->          |←———180———→|
|←15-16→|
```

Обозначения :

----> обычный повтор

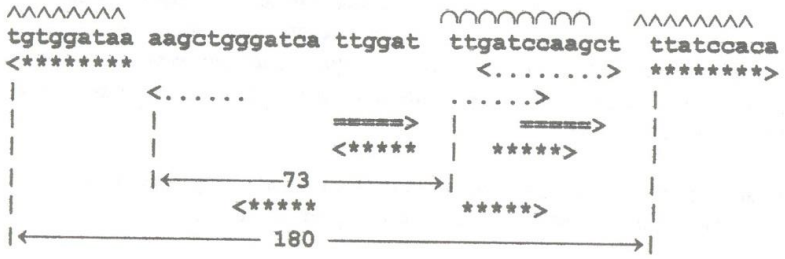
******> комплементарный повтор**

Рис. 2 Симметрийные консенсусы для последовательностей {ec,ea,kp,er,sd,st,vh}

Группа {ec, ea, kp, er, sd, st, vh} имеет симметрийный консенсус, состоящий из одного комплементарного повтора - ttatccaca (рис 2, а), представляющего собой сайт связывания белка DnaA. При удалении из

группы последовательности {vh} симметрийный консенсус обогащается вторым комплементарным повтором - agatct, который в силу внутренней симметрии является одновременно и прямым обычным повтором (Рис2, б).

При удалении из группы {ec, ea, kp, er, sd, st} последовательности {er}, симметрийный консенсус существенно обогащается в том числе и структурами симметрии RY и KM (Рис.3). На рисунке 3 представлена часть симметрийного консенсуса для группы {ec, ea, kp, sd, st}. Исключение из группы E.coli {ec, ea, kp, sd, st} любой из пяти последовательностей не приводит ни к значительному увеличению общей длины словаря симметрийного консенсуса, ни к его обогащению новыми типами симметрии. Из восьми типов симметричных структур в симметрийном консенсусе группы {ec, ea, kp, sd, st} не входят только инвертированные RY и прямые KM повторы.



Обозначения:

∩∩∩ сайт узнавания ДНКазы I

∧∧∧∧сайт связывания DnaA белка

****> комплементарные повторы

.....> KM - повторы

=====> RY- повторы

Рис.3 Часть симметрийного консенсуса для группы последовательностей {ec, ea, kp, sd, st}.

Сравнение структуры построенных в работе симметрийных консенсусов с экспериментальными данными о строении ori показало, что симметрийные консенсусы, основанные на наличии определенных симметричных структур, элементы которых расположены на равных расстояниях, включают все типы экспериментально выявленных структурных элементов ori прокариот, а именно:

1). сайты взаимодействия с белками (в состав СК входят DnaA-боксы R1 и R4 и часть DnaA-бокса R3, а также сайты взаимодействия с негативным регулятором инициации репликации SeqA белком);

2). ДНК-расплетающий элемент DUE (СК содержат части левого L и среднего M 13-меров);

3). все выявленные сайты разделены фиксированными расстояниями, которые являются спейсерами.

Наличие в построенных нами симметрийных консенсусах именно этих экспериментально выявленных структурных элементов ori позволяет предположить, что симметрийный консенсус представляет структуру в генетическом тексте прокариотических сайтов инициации репликации, узнаваемую инициирующими белками на самых начальных стадиях инициации репликации.

Симметрийный консенсус для последовательностей рода Pseudomonas

На рис.5 показан симметрийный консенсус, построенный для двух последовательностей из рода Pseudomonas, который также состоит из двух блоков. Разница в расстояниях между блоками у {pa} и {pp} равна 70 нуклеотидам, что соответствует семи полным виткам спирали ДНК, т.е. блоки находятся на определенной стороне спирали и строго ориентированы друг относительно друга в пространстве.

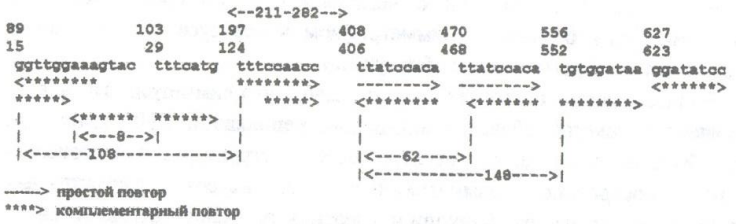


Рис. 5. СК для сайта инициации репликации рода Pseudomonas

В связи с этим хотелось бы отметить, что хотя для последовательностей {pl} и {sc} симметрийные консенсусы не строились, но комплементарные повторы, соответствующие сайтам связывания DnaA белка, расположены в {sc} на расстоянии 389 нуклеотидов, а в {pl} - на расстоянии 116 нуклеотидов. Разница в расстояниях между DnaA-боксами в симметрийном консенсусе для E.coli и в {sc} соответствует 21 полному витку спирали, а разница в расстояниях между DnaA-боксами в СК E.coli и в последовательности {pl} - 10 полных витков спирали ДНК. Таким образом, проведенный симметрийный анализ исследуемых нуклеотидных последовательностей, обладающих функцией ori, показал, что в изученных прокариотических ori наблюдается одинаковая ориентация структурных

блоков относительно друг друга, что, вероятно, является важным принципом пространственной организации сайтов инициации репликации прокариот.

Сравнение построенных в работе симметрийных консенсусов двух видов, для группы E.coli и для Pseudomonas, показало следующее:

- И в том , и в другом случае симметрийные консенсусы состоят из двух блоков: в правый входят сайты связывания с DnaA белком, слева расположен другой блок, образованный обычными и комплементарными повторами.
- Расстояние между блоками фиксировано (40- 42 нуклеотида) для последовательностей группы E.coli. В последовательностях рода Pseudomonas блоки расположены на разных расстояниях (211 нуклеотидов в последовательности {pa} и 282 нуклеотида в {pp}), но одинаково ориентированы относительно друг друга и оси спирали ДНК
- В симметрийных консенсусах для последовательностей группы E.coli в правом блоке обнаружены симметричные структуры RY- и KM-типов, которые выявляются во всех последовательностях этой группы, что указывает на их значимость для функции инициации репликации ДНК.
- В симметрийных консенсусах группы последовательностей E.coli выявлены 4 сайта связывания с негативным регулятором инициации репликации SeqA белком. В симметрийном консенсусе Pseudomonas сайтов негативной регуляции не обнаружено.

Вышесказанное позволяет сделать вывод, что симметрийный анализ генетических текстов областей инициации репликации ДНК прокариот может быть использован для исследования их структуры, что позволяет считать построение симметрийных консенсусов эффективным инструментом для анализа функции нуклеотидных последовательностей.

Использование симметриных консенсусов для распознавания ori.

Из анализа экспериментальных данных можно заключить, что сайт инициации репликации ДНК прокариот представляет собой структуру со сложным внутренним строением. Его можно представить как список слов в алфавите {a, c, g, t} с их позициями относительно некоторой точки и искать его именно в таком виде. Это означает, что функциональное ядро ori может быть представлено как множество слов, связанных определенными отношениями типа «содержит», «слева», «длина», «расстояние», «позиция».

«Экспериментальный» образ (OriC)

Для описания oriC в качестве словаря использовались экспериментально выявленные сайты, полученные из различных

литературных источников (последовательности рассматриваются в принятой ориентации 5' - 3'):

S_0=oriC
S_1=DnaA-boxR1 TTATCCACA
S_2=DnaA-boxR2 TTATACACA
S_3=DnaA-boxR3 TTATCCAAA
S_4=DnaA-boxR4 TTATCCACA
S_5=DnaA-boxM TCATTCACA
S_6=13merR GATCTATTTATTT
S_7=13merM GATCTGTTCTGTT
S_8=13merL GATCTCTTATTAG
S_9= AT- rich
S_{10}= DUE

Следующие отношения могут описывать пространственные отношения между этими подстроками:

Отношение	Тип	Смысл
длина (s)= N	целое число	число нуклеотидов
содержит(s_1, s_2)	логический	s1 полностью содержится в s2
позиция (s)	целое число	позиция 5' конца
расстояние (s_1,s_2)	целое число	Позиция (s3,s2) - Позиция (s1,s2)
слева (s_1,s_2)	логический	1 находится слева от s2

Используя эти отношения, oriC можно представить как набор следующих утверждений:

длина (ori) = 240
длина ($S_{1\text{-}5}$) = 9
длина ($S_{6\text{-}8}$) = 13
содержит (ori, $S_{1\text{-}10}$) = да
содержит (S_9, S_{10}) = да
содержит (S_9, S_6) = да
содержит (S_9, S_7) = да
содержит (S_9, S_8) = да
позиция (ori) = 0
расстояние (S_1,S_4)= 180
расстояние (S_1,S_2)= 95
расстояние(S_2,S_3)= 27
расстояние(S_3,S_4)= 30
расстояние(S_1,S_5)= 46
расстояние (S_5,S_2)= 27
слева (S_9, S_1)

слева (S_1, S_5)
слева(S_5, S_2)
слева (S_2, S_3)
слева (S_3, S_4)
слева (S_6, S_1)
слева(S_7, S_1)
слева (S_8, S_1)

«Экспериментальный» образ является корректным только для oriC E.coli, попытка применения этой модели для описания других прокариотических ori, даже близких к oriC E.coli, приводит к увеличению возможных вариантов этой модели, что сразу усложняет задачу распознавания. Эту проблему можно преодолеть, если пространственные отношения дополнить отношениями преобразования элементов словаря друг в друга.

«Формальный» образ ori

«Формальный» образ строился с использованием тех же отношений, которыми пользовались для описания «экспериментального» образа, но в качестве элементов словаря использовались симметричные структуры, обнаруженные в исследованных последовательностях. Обозначения для симметричных преобразований приведены в таблице 4.

Таблица 4 Обозначение симметричных преобразований элементов словаря.

Типы повторов	Прямые	Инвертированные
Простые	P	iP
Комплементарные	P^C	iP^C
RY	P^R	iP^R
KM	P^K	iP^K

Использование обозначений симметричных преобразований и описанных выше предикатов позволяет описать «формальный» образ ori для группы последовательностей {ec, ea, kp, er, sd, st}, представленный на рис. 2 , в виде следующих утверждений :

содержит (ori, S_1)
содержит (ori, $iP^C(S_1)$)
слева ($iP^C(S_1)$, S_1)
расстояние (S_1, $iP^C(S_1)$) = - 180 или расстояние ($iP^C(S_1)$, S_1) = 180
содержит (ori, S_2)

содержит (ori, $P(S_2)$)
расстояние (S_2, $P(S_2)$) =15-16
слева (S_2, $iP^C(S_1)$)
расстояние (S_2, $iP^C(S_1)$) = 40-42, где S_1 = ttatccaca
S_2 = agatct

Для симметриного консенсуса, представленного на рисунках 2,б и 3, можно предложить следующее описание:
Словарь: S_1 = ttatccaca
S_2 = agatct
S_3 = ccaagc
S_3 = ttgatcc
содержит (ori, S_1)
содержит (ori, $iP^C(S_1)$)
слева ($iP^C(S_1)$, S_1)
расстояние ($iP^C(S_1)$, S_1) = 180
содержит (ori, S_2)
содержит (ori, $P(S_2)$)
расстояние (S_2, $P(S_2)$) = 15-16
слева (S_2, $iP^C(S_1)$)
расстояние (S_2, $iP^C(S_1)$) = 40-42
содержит (ori, S_3)
содержит (ori, $P^R(S_3)$)
слева ($P^R(S_3)$, S_3)
расстояние ($P^R(S_3)$, S_3) = 20
содержит (ori, S_4)
содержит (ori, $iP^K(S_4)$)
слева ($iP^K(S_4)$, S_4)
расстояние ($iP^K(S_4)$, S_4) = 73
слева (S_3, S_1)
слева (S_2, S_3)
слева (S_4, S_1)
слева (S_2, S_4)

«Формальный» образ сайта инициации репликации рода Pseudomonas, симметрийный консенсус для которого представлен на рис. 5, описываеся следующим образом:
Словарь: S_1 = ttatccaca
S_2 = tttccaacc
содержит (ori, S_1)
содержит (ori, $P(S_1)$)

слева (S_1, $P(S_1)$)
расстояние (S_1, $P(S_1)$) = 62
содержит (ori, $iP^C(S_1)$)
слева (S_1, $iP^C(S_1)$)
расстояние (S_1, $iP^C(S_1)$) = 148
содержит (ori, S_2)
содержит (ori, $iP^C(S_2)$)
слева ($iP^C(S_1)$, S_1)
расстояние ($iP^C(S_1$, S_1) = 108
слева (S_2, S_1)
расстояние(S_2, S_1) = 211 или 282

Таким образом, использовав в качестве элементов словаря обнаруженные в исследуемых последовательностях симметричные структуры, и, связав эти элементы определенными выше отношениями, мы получили компактное формальное описание ori прокариот.

«Формальный» образ ori, основанный на симметриях, в достаточной степени формализован для компьютерного представления и его можно использовать для распознавания ori, для поиска и идентификации сайтов инициации репликации ДНК в генетических текстах прокариот.

Формальная модель ori была использована для поиска симметрийного консенсуса, определенного для группы E.coli, в известных последовательностях ДНК кишечной палочки в банке данных нуклеотидных последовательностей EMBL (Genomes, E.coli), который показал, что полученный нами симметрийный консенсус для последовательностей группы E.coli обнаруживается только в единственном месте хромосомы E.coli - сайте инициации репликации oriC. Это означает, что предложенная в работе модель ori в виде множества симметричных структур с указанием их позиций пригодна для распознавания ori в геноме.

ВЫВОДЫ:

1. Осуществлен поиск симметричных структур и выявлены повторы всех возможных 8-ми типов симметрии в нуклеотидных последовательностях 13-ти исследованных сайтов инициации репликации прокариот.
2. Были выявлены симметричные структуры в генетических текстах исследованных прокариотических ori, значимые для функции инициации репликации. Было показано, что важными для функции инициации репликации наряду с традиционно рассматриваемыми обычными и комплементарными повторами оказались не рассматривавшиеся ранее симметричные структуры RY и KM типов.

3. Проведена классификация исследуемых последовательностей по признаку вхождения определенных симметричных структур, которая хорошо согласуется с таксономией организмов, которым принадлежат изученные последовательности. Обнаружено также точное соответствие формальной классификации исследованных последовательностей и классификации прокариотических ori, основанной на экспериментальных данных, что является подтверждением функциональной значимости выявленных симметричных структур для инициации репликации.
4. Построены симметрийные консенсусы (СК), основанные на наличии в последовательностях симметричных структур, элементы которых расположены на одинаковых расстояниях. Показано, что построение симметрийных консенсусов для генетических текстов сайтов инициации репликации прокариот является эффективным инструментом для изучения внутренней структуры этих сайтов.
5. В работе показано, что симметрийный консенсус, построенный для группы последовательностей E.coli, обнаруживается в единственном месте хромосомы - в сайте инициации репликации oriC. Это означает пригодность предложенных в работе моделей сайта инициации репликации ДНК для распознавания ori в геноме прокариот.

Публикации по теме диссертации:

1. Akberova N.I., Leontyev A.Yu. Symmetry analysis of the genetic text in gene identification and phylogenetic analysis. Genome. Science & Technology, 1995, v.1, No 1, p. P-60, C-1

2. Akberova N.I., Leontyev A.Yu.The symmetry patterns' analysis as a tool for recognition of functional sites and phylogenetic trees construction. BIOMATH-95, DATECS publishing, Sofia. 1995, p.21

3. Akberova N.I., Leontyev A.Yu. Prokaryotic replication initiation protein complex may recognize symmetry structures in the DNA genetic texts of replication origins The Proceedings of the Protein Society Ninth Symposium (July 8-12, 1995, Boston, MA), 382-T, B197

4. Akberova N.I., Leontyev A.Yu. Functional site identification by means of the symmetry analysis of the genetic text. Abstracts of the 23th Meeting of the Federation of European Biochemical Societies (FEBS'95, Basel, Switzerland, 13- 18 Aug 1995)

5. Akberova N., Leontyev A.Yu. Functional site identification and phylogenetic analysis by means of genetic text symmetry analysis method. The Book of Abstracts of the International Conference on Molecular Structural Biology (Vienna, Austria, September 17-20 1995)

6. Akberova N.I., Leontyev A.Yu. Functional site identification and phylogenetic analysis by means of genetic text symmetry analysis method.Karadeniz Journal of Medical Sciences 1995, v.8, No.4, p.209-210

7. Akberova N.I., Leontiev A.Yu. Symmetrical structures in genetic texts of prokaryotes DNA replication origins NetSci 1996, v.2, March on-line http://www.awod.com/netsci/Issues/Oct95 /feature4.html

8. Akberova N.I., Leontyev A.Yu. A context-free grammar of the DNA functional sites, based on symmetrical structures in the genetic text. The GCB`96 Proceedings, 1996. (German Conference in Bioinformatics , Leipzig, September 30 - October 2, 1996)

9. N.Akberova, A.Leontiev, V.Vinter The formal analysis of the DNA sequences presented as a text in the alphabet based on the symmetry of the genetic text Abstracts of PSB97(Pacific Symposium on Bioinformatics, Hawaii, January, 1997)

10. Akberova N.I., Leontyev A.Yu. Application of symmetrical structures in the genetic text for DNA functional sites' grammar construction. The ISMB-97 Proceedings (The 5th International Conference on Intellegent Systems for Molecular Biology, Halkidiki, Greece, June 21-25, 1997)

11. Akberova N.I., Leontyev A.Yu. Application of Symmetry Patterns to Recognition of Functional Sites and Phylogenetic Analysis of DNA Sequenses. The Proceedings of SIS'98(the IEEE Symposia on Intelligence and Systems - (SIS'98), the Intelligence in Neural and Biological Systems conference (INBS). May 21-23, 1998, Washington DC, USA. Copyright 1998 IEEE.

12. N.Akberova, A.Leontiev. The Symmetrical Analysis of Genetic Texts of the Prokaryotic Replication Initiation Regions Proceedings of ISMB98 Workshop "Semantic foundations for molecular biology" (6th International Conference on ISMB, Montreal, June 28 - July 1, 1998)

13. Akberova N.I., Kolpakov A., Leontiev A.Yu. Context-free method of pattern recognition in the genetic texts Proceedings of the First International Conference of Bioinformatics of Genome Regulation and Structure (BGRS'98) Novosibirsk-Altai Mountains, Russia, August 24-31, 1998, published by ICG, Novosibirsk, 1998, vol. , pp

Подписано в печать 19.03.99.
Усл.печ.лист. 1.5. Тираж 50. Договор № 5.
Отпечатано в лаборатории оперативной полиграфии ТГГИ
г. Казань, ул Побежимова 47а., Тел 54-43-73.